DOCUMENT RESUME

ED 128 381 TM 005 503

AUTHOR Durovic, Jerry J.
TITLE Test Bias: An Objective Definition for Test Items.
PUB DATE [Oct 75]
NOTE 17p.; Paper presented at the Annual Meeting of the Northeastern Educational Research Association (Ellenville, New York, October 1975)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Content Analysis; *Definitions; *Item Analysis; Mathematical Models; Probability; *Test Bias
IDENTIFIERS Rasch Model

ABSTRACT

A test bias definition, applicable at the item-level of a test is presented. The definition conceptually equates test bias with measuring different things in different groups, and operationally equates test bias with a difference in item fit to the Rasch Model, greater than one, between groups. It is suggested that the proposed definition avoids confusing etiology with measurement by capitalizing on the "objectivity" property of the logistic Rasch Measurement Model. Application of the definition, to 914 applicants (black = 367; white = 547) in a "real" selection situation is described and resulted in identifying two items as biased. The two items so defined, were different than the two items identified as biased by comparing the item success rates (i.e., item difficulty) of black vs. white candidates. A content evaluation of the items by two black, female reviewers was subsequently performed. Their comments lend preliminary support to the proposed psychometric test bias definition. Additional encouraging support is provided by the match between the content comments and the item bias index values, for other items in the test. Implications for future applications and research are presented. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED128381

TEST BIAS: AN OBJECTIVE DEFINITION FOR TEST ITEMS

Jerry J. Durovic

New York State Department of Civil Service

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM005 503

Paper presented as part of Symposium entitled: "Test Bias: Some Conflicting Unbiased Views," 1975 Annual Convocation of the Northeastern Educational Research Association, Ellenville, October 30, 1975.

Psychometric definitions of bias vary on three dimensions, each a dichotomy, for a total of eight possible cells in the classification scheme (Durovic, 1975). These dimensions include the score-level addressed (i.e., item vs. test), the role of an external criterion variable, and the role of external value judgments. The dimension important to the present paper is the score-level addressed; only definitions at the item level are considered. The present paper is concerned with one of the eight possible cells: item-level definitions with no external criterion variable and no external value judgments needed.

Test Bias Definitions at the Item Level

Few test bias definitions are addressed to the item-level. The simplest definition, at the total test level, equates test bias with group differences in test score means. For the item-level, it can be modified to equate bias with group differences in item score means. Test constructors have traditionally used this approach to eliminate those group differences they consider irrelevant to the purposes of the test. For example, items for the Stanford-Binet were chosen, in part, to eliminate score differences between sexes (Anastasi, 1968). However, this approach to test bias is faulty, since:

> tests whose items have been selected with reference to the responses of any special groups cannot be used to compare such groups (Anastasi, 1968, p. 179).

Critics of the mean-difference definition argue that, if the test accurately identifies existing differences in the ability being measured, then differential group performance should not label a test as biased. To do otherwise confuses etiology with measurement (Anastasi, 1961). Conceptually, test bias applies to the properities of the instrument and not to the persons taking it (Durovic,1975). The problem of etiology vs. measurement exists whether the mean-difference

definition is applied to the item or to the total test score level.

Another definition of test bias, at the item-level, is provided by Cleary and Hilton (1968). They state that an item is

> biased for members of a particular group, if on that item, the members of the group obtain an average score which differs from the average score of other groups by more or less than expected from performance on other items of the same test. That is, the biased item produces an uncommon discrepancy between the performance of members of other groups. In terms of the analysis of variance, bias is defined as an item X group interaction (Cleary and Hilton, 1968, p.61).

The Cleary-Hilton definition of test bias has been used by other researchers such as Angoff and Ford (1971), Echternacht (1972), and Hoepfner and Strickland (1972). A variant of this definition was employed by Evans and Reilly (1972) who investigated speededness as a source of test bias and defined bias as the time x race interaction. In all of these studies the ANOVA method is used and the interaction term with race serves as the definition of test bias.

This definition differs from the simple mean difference definition. However, the ANOVA approach still equates mean item score differences with test bias; it simply allows the difference to be determined by the group performance on the total set of items, rather than mandating no allowable difference between groups.

Objective Test Bias Definition

An approach to defining test bias at the item level, which avoids the etiology vs. measurement problem, may be provided by the Rasch Model

(Rasch, 1960). Georg Rasch developed several measurement models and one of these models, referred to here simply as the Rasch Model, applies to test items which can be viewed as a dichotomous, or binomial response situation of correct vs. incorrect response (Rasch, 1960; 1966a; 1966b). The Rasch Model involves two parameters: one parameter for the person's ability level, and one parameter for test item easiness level. Some researchers refer to the Rasch Model as a one parameter model because it has only one parameter for test items (Hambleton and Traub, 1970).

The attractiveness of the Rasch Model as an approach to defining test bias is due to its property of "objectivity" (Tinsely and Dawis, 1975; Wright, 1968) which makes it possible, in the analysis of data, to separate the person parameter from the test item parameter (Rasch, 1961). Essentially the Rasch Model "objectivity" property permits calibration of test items independent of the sample distribution of ability in the subjects. Thus, even if one group of subjects is dramatically different from a second group of subjects on the ability being measured, the same item calibrations should occur for each group separately. That is, "objectivity"

> requires that test calibration be independent of which persons are used for the calibration and that person measurement be independent of which items are used for the measurement (Wright, 1968, p. 1).

Therefore, it appears possible to avcid, mathematically, the problem of confusing person ability levels with test instrument characteristics, if we can capitalize on the property of "objectivity."

One approach to defining bias in this way, at the item level, is presented here. During calibration an independent test of fit to the Rasch Model is available for each item. Since, the Model is a probabilistic one, a

perfect fit of data to the Model is not expected. A deviation of approximately one standard deviation for the observed data from the expected is projected, that is, a mean square fit of an item to the Model equal to one is expected (Wright, Note 1). Therefore, test bias is defined here as follows: an item is biased for members of a group, if on that item, for members of the group, a mean square fit of the item to the Rasch Model is obtained which differs, by greater than one, from the mean square fit obtained for members of the other group. By this definition, a test is not biased if each item in the test relates to the dimension being measured in the same way for each group.

Hypothetical Illustration

Essentially, in a measurement situation we have two groups of elements. One group consists of the measuring instruments, and the second group consists of the objects to be measured. Typically, the first group of instruments are test items, and the second group of objects are persons. For illustrative purposes, let us consider for the moment, that the first group consists of instruments that exert a reproducible influence readily acceptable as mechanical only (i.e., they "push"), and the second group consists of objects readily acceptable as solid bodies. Then if each instrument, I_j, is applied to each object, O_r, a set of rates of accelerations, A_{rj}, may be observed.

Then if it happened that the acceleration received by O_1 from instrument I, is twice that received by O_2 from the same instrument, and it was further found, aside from errors in measuring the acceleration, the same sort of results when applying any other instrument then it could be said O_1 has half the "mass" of O_2. Note, the mean-difference definition would label such an instrument, I, as biased.

Next, against this background suppose the mistake was made of not noticing that one of the instruments was strongly magnetic. If among the solid

bodies were both a piece of wood and a piece of steel then obviously magnetism would play havoc with the accelerations. The accelerations produced by this instrument would not fit in with the mechanically produced accelerations. If we did not make too many such mistakes then we would easily discover and locate the error. Within the context of the "acceleration" experiment, the instruments which discriminate between the metal and non-metal objects (i.e., the magnetic instruments) are "biased" by the definitional approach presented here.

Empirical Illustration

To empirically explore the definition offered here, selection-test responses of 914 adult candidates (Black 367; White 547), for a wide variety of client-oriented government positions, were analyzed. A 14 item multiple-choice test, that was part of a written test for the positions, was used as the test instrument.

The subjects were identified by race based on a self-administered questionnaire. A total of 367 candidates identified themselves as Black and 547 candidates identified themselves as White, at those test centers with self-identified Black candidates.

Applying the Cleary-Hilton (1968) definition (see Table 1) resulted in finding a significant interaction effect of race x items ($p \leq .01$) and therefore by this definition there is test bias (see Table 2). The items with the largest item difficulty differences are items #11 and #15 (see Table 3).

Applying the proposed definition resulted in finding two items, #8 and #13, with a mean square fit $\hat{\Delta}$ greater than one (see Table 4). Therefore, by this definition, only items #8 and #13 are biased. These two items are not the same as the two items with the greatest difference between groups in average score.

Content Evaluation

Next, the possibility of determining an item-content-based explanation for the difference between groups in fit of an item to the Rasch Model (i.e., the proposed test bias item index) was considered. There were several reasons for believing that content explanations might be possible. First, Rasch (1960) suggests this possibility when he states:

> once a law has been established within a certain field then the law itself may serve as a tool for deciding whether or not added stimule and/or objects belong to the original group (p. 124).

Second, in discussing the Rasch Model, Wright and Panchapakesan (1969) state that a

> source of lack of fit of an item lies in the content of the item. The Model assumes that all items used are measuring the same trait. Items in a 'test' may not fit together if the 'test' is composed of items which measure different abilities. This includes the situation in which the item is so badly constructed or so mis-scored that what it measures is irrelevant to the rest of the 'test' (p. 25).

Third, empirical support for content-based explanations of misfitting items has been presented. In a series of studies, using personnel selection test responses of adult applicants, Durovic (1970) reported that the misfitting items had test construction defects or were measuring different abilities from those of the total test. More recently, Kifer and Bramble (1974) reported that an examination of the items which did not fit the Rasch Model "indicated that about one-half were poorly written" (p. 2).

Therefore, the feasibility of revealing possible content-based support for the items deemed biased (i.e., items #8 and #13) by the proposed item bias index was explored.

Two reviewers were selected who had no role in the preparation of the test used in this study and, therefore, could provide and independent content evaluation. These two reviewers were selected because it was felt they could evaluate the tests for possible bias against Blacks.

The first reviewer was a Black female member of the New York State Advisory Committee to the United States Commission on Civil Rights and is involved in assisting the Committee to determine the extent of discrimination in selection practices in public employment.

Her reaction to the two items identified as biased by the proposed item bias index was immediate and intense. She had a strong, emotional, negative reaction to them, and argued with the premise of the items. This reaction did not exist in the remaining twelve items.

The second reviewer was a Black female official of the International Personnel Management Association (Eastern Region) and Director of a large test development group that constructs public sector employment tests for government jurisdictions in New York State. She submitted written comments on each item in a narrative format (except for item #6, which she presented in an outline format, and as a result item #6 is not considered here further) which ranged from a succinct "seems o.k." (Griffin, Note 2) for items #10 and #12, to lengthy detailed criticism.

First, her reaction to the two items identified as biased by the proposed definition were in substantial agreement with the first reviewer. For item #8, she states the item would have an emotional impact on Blacks. For item # 13, she argues with the item and states the key answers are not appropriate or even correct for most minorities. Second, she gave twice as

many lines of criticism to two items in the test than to any other items (i.e., nine lines vs. five lines). The two items which received the most criticism were, items #8 and #13, the two defined as biased by the proposed psychometric definition. Third, only two items received five (5) lines of criticism (i.e., item #1 and item #7). Interestingly, they both received virtually identical test bias item indices (i.e., mean square fit difference values) by the proposed definition. Fourth, three, and only three, items received criticism that the vocabulary or terms used in these items was not appropriate for minority group members (i.e., items #2, #5, and #6). The test bias item index, for these three items , cluster together and no other items have a similar index.

While no firm conclusions can be placed on these two reviews, their comments lend some preliminary content support to the psychometric test bias decision reached by the proposed definition. In addition, the observation that some items, with similar content comments from the second reviewer had similar test bias item indices, suggests that perhaps item index differences themselves may be fruitful areas of study (i.e., analogous to discovering magnetic bodies).

Conclusion

A test bias definition, applicable at the item-level of a test is presented. The definition conceptually equates test bias with measuring different things in different groups, and operationally equates test bias with a difference in item fit to the Rasch Model, greater than one, between groups. It is suggested that the proposed definition avoids confusing etiology with measurement by capitalizing upon the "objectivity" property of the Rasch Model. Application of the definition, to applicants in a "real" selection situation, resulted in identifying two items as biased. The two items, so defined, were different than the two items identified as biased by comparing differential item success rates (i.e., item difficulty). A content evaluation of the items by two Black,

female reviewers was subsequently performed. While no firm conclusions can be placed on these two reviews, the comments lend preliminary support to the proposed psychometric test bias definition. Additional encouraging support is provided by the match between the content comments and the item bias index values, for other items in the test.

Future research might consider the applicability of the definition proposed here as a procedure to refine tests, identify possible general reasons for test bias (e.g., vocabulary problems), and examine the impact on criterion-related validity. For example, in the test evaluated here, some item clustering by the psychometric procedures of the proposed test bias definition appear to parallel the item clustering by the subjective content evaluations of the minority group reviewers. Strong emotional reaction and differential vocabulary useage were suggested as possible explanations for two of the item clusters. Can these findings be replicated or generalized? If replicable and generalizable content-related explanations for bias can be found for items, psychometrically defined as biased by the proposed definition, then several practical implications follow. First, test developers could refine their tests by constructing substitute items for the biased ones and then psychometrically evaluate their effort. This test refinement process, when coupled with standard validation procedures may improve tests used for selection. Second, educators could consider setting goals to eliminate the content-related disparity identified by the biased items. Evaluators might psychometrically assess the success in achieving the goals, by administering the biased items and applying the proposed definition. It should be noted that this procedure is not the same as evaluating item success rates or achievement. If the reason of the bias has been removed then the test bias index should reflect it, within the context described earlier. Thus, if biased items are found, then, either they can be removed if subject discriminations on this dimension are undesired, or, they can be used exclusively if subject discriminations on this dimension is desired.

References

Anastasi, A. "Psychological tests: Uses and Abuses." Teachers College Record, 1961, 62, 389-393.

Anastasi, A. Psychological Testing, New York: Macmillan Company, 1968.

Angoff, W. and Ford, S. "Item-race interaction on a test of scholastic aptitude." Research in Education, 1972, 1-26, TM000992.

Cleary, T. and Hilton, T. "An investigation of item bias." Educational and Psychological Measurement, 1968, 28, 61-75.

Durovic, J. "Application of the Rasch Model to civil service testing." Paper presented to the Northeastern Educational Research Association, Grossingers, New York, 1970, ED 049305.

Durovic, J. "Definitions of test bias: A taxonomy and an illustration of an alternate model." Unpublished doctoral dissertation, State University of New York at Albany, 1975.

Echternacht, G. "A quick method for determining test bias." Educational and Psychological Measurement, 1972, 14, 271-280.

Evans, F. and Reilly, R. "A study of speededness as a source of test bias." Journal of Educational Measurement, 1972, 2, 123-131.

Hambleton, R. and Traub, R. "Information curves and efficiency of three logistic test models." Technical Report No. 4, Center for Educational Research, School of Education, University of Massachusetts, Amherst, Massachusetts, 1970.

Hoepfner, R. and Strickland, G. "Investigating test bias." Research in Education, 1972, 1-35, TM001772.

Kifer, E. and Bramble, W. "The calibration of a criterion-referenced test." Paper presented at American Educational Research Association, Chicago, Illinois, 1974.

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: The Danish Institute of Educational Research, 1960.

Rasch, G. "On General Laws and the Meaning of Measurement in Psychology." Proceedings of the Fourth Berkely Symposium on Mathematical Statistics. Berkeley: University of California Press, 1961, IV, 321-334.

Rasch, G. "An Individualistic Approach to Item Analysis." In Readings in Mathematical Social Science. Edited by Lazarsfeld and Henry. Chicago: Science Research Associates, Inc. 1966a, 89-107.

Rasch, G. "An item analysis which takes individual differences into account." British Journal of Mathematical and Statistical Psychology. 1966b, 19, 49-57.

Tinsley, H. and Dawis, R. "An Investigation of the Rasch Simple Logistic Model: Sample Free Item and Test Calibration." Educational and Psychological Measurement, 1975, 35, 325-339.

Wright, B. "Sample-free test calibration and person measurement." In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton: Educational Testing Service, 1968, III, 85-101.

Wright, B. and Panchapakesan, N. "A procedure for sample-free item analysis." Educational and Psychological Measurement, 1969, 29, 23-48.

Reference Notes

1. Wright, B. D. Personal communication, 1969.

2. Griffin, M. G. Personal communication, April 23, 1975.

Table 1

Cleary-Hilton ANOVA Model for Data

$$X_{is(r)} = M + I_i + R_r + S_{s(r)} + IR_{ir} + IS_{is(r)}$$

Where M = grand mean

I_i = items ($i = 1,\ldots,14$)

R_r = race ($r = 1,2$; 1 = black, 2 = white)

$S_{s(r)}$ = subjects within race [$s_{(r=1)} = 367$; $s_{(r=2)} = 547$]

Table 2

Cleary-Hilton ANOVA Test Bias Definition

Source	df	MS	F
Item (I)	13	23.28	139.99*
Race (R)	1	2.72	9.77*
Subjects (S)	912	.28	
I x R	13	2.03	12.15*
I x S	11,856	.17	

* $P \leq .01$

Table 3

Test Item Difficulty for Each Racial Group

Item[a]	Race				
	Black (B)		White (W)		(B-W) Difference
	Count	Percent	Count	Percent	Percent
1	280	77.56	473	88.08	-10.52
2	265	73.41	405	75.42	- 2.01
3	113	31.30	229	42.64	-11.34
4	300	83.10	467	86.96	- 3.86
5	247	68.42	364	67.78	0.64
6	212	58.73	372	69.27	-10.54
7	208	57.62	331	61.64	- 4.02
8	239	66.20	424	78.96	-12.76
9	201	55.68	282	52.51	3.17
11**	342	94.74	528	58.32	36.42
12	338	93.63	519	96.65	- 3.02
13	213	59.00	301	56.05	2.95
14	253	70.08	440	81.94	-11.86
15*	327	90.58	351	65.36	25.22

[a]item #10 deleted
*item with second largest item difficulty difference
**item with largest item difficulty difference

Note-A constant should be subtracted from the (B-W) Difference shown in order to evaluate in terms of the Cleary-Hilton Definition, however in the present instance since $M_1 - M_2 < 0.5$, then 0.5/14 makes little difference.

Table 4

Mean Square Fit of Items to Rasch Model for Both Groups

Item[a]	Group		
	Black (B)	White (W)	(B-W) Difference
1	1.03	0.99	0.04
2	1.87	1.02	0.85
3	0.79	1.41	-0.62
4	0.88	1.01	-0.13
5	1.85	1.09	0.76
6	1.16	0.55	0.61
7	0.83	0.76	0.07
8	1.13	2.20	-1.07*
9	1.29	1.48	-0.19
11	0.62	0.43	0.19
12	0.88	0.63	0.25
13	2.16	0.87	1.29*
14	1.11	0.66	0.45
15	1.51	1.28	0.23
All	1.31	1.11	0.20
Chi-square probability	0.009	0.206	
df	130	117	

[a]item #10 deleted
*Difference > 1.00